Can a Virtual Human Facilitate Language Learning in a Young Baby?

ABSTRACT

The child developmental period of ages 6-12 months marks a widely understood "critical period" for healthy language learning, during which, failure to receive exposure to language can place babies at risk for language and reading problems spanning life. Deaf babies constitute one vulnerable population as they can experience dramatically reduced or no access to usable linguistic input during this period. Technology has been used to augment linguistic input (e.g., auditory devices; language videotapes) but research finds limitations in learning. We evaluated an AI system that uses an Avatar (provides language and socially contingent interactions) and a robot (aids attention to the Avatar) to facilitate infants' (6-13 months) ability to learn aspects of American Sign Language (ASL) produced by our signing Avatar, and asked two questions: (1) Can babies with little/no exposure to ASL distinguish among the Avatar's different conversational modes (Linguistic Nursery Rhymes; Social Gestures; Idle/nonlinguistic postures; 3rd person observer)? (2) Can an Avatar stimulate babies' production of socially contingent responses, and crucially, nascent language responses? (3) What is the impact of parents' presence/absence of conversational participation? Surprisingly, babies (i) spontaneously distinguished among Avatar conversational modes, (ii) produced varied socially contingent responses to Avatar's modes, and (iii) parents influenced an increase in babies' response tokens to some Avatar modes, but the overall categories and pattern of babies' behavioral responses remained proportionately similar or the same irrespective of parental participation. Noteworthy, babies produced the greatest percentage of linguistic responses to the Avatar's Linguistic Nursery Rhymes versus other Avatar conversational modes. This work demonstrates the potential for Avatars to facilitate language learning in young babies.

KEYWORDS

Empirical studies on social agents/robots; Social impact; Multi-user/multi-agent/robot interaction

1 INTRODUCTION

Many AI systems have been designed for facilitating language learning by adults, and to a lesser extent, children [4, 9, 29]. However, there is a significant paucity of work on AI systems for young infants despite the widely understood critical importance that this developmental period has for healthy language and cognitive growth, and related reading and academic success [22]. Children have proven to be a challenging population to design language learning technology, and some technologies designed for children have been shown not to facilitate language learning. For example, Krcmar [11], Krcmar et al. [12], Richert et al. [25] have described studies showing that children who receive linguistic stimuli from television do not learn as much as those who receive the same linguistic stimuli from live adults. Our particular interest is young babies who lack the necessary language exposure in early life. Based on the discovery of brain-based "critical periods" of human development for language (e.g., ages 6-12 months; [19]), a growing body of neuroscience research has identified the potentially devastating impact that minimal language exposure during this particular period of child development can have on all children's linguistic, cognitive, and social skills, be they hearing or deaf infants [18]. As such, technology that can help fill a language-exposure gap can have a tremendous impact for social good in these populations. Deaf babies constitute one especially vulnerable population, as they can experience dramatically reduced or no access to usable linguistic input during this period.

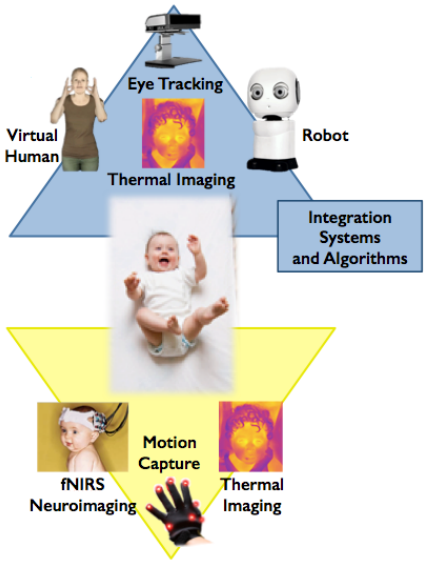


Figure 1: The RAVE AI Language Learning System+ Human Infant

One recently introduced AI system, called RAVE (Robot, AVatar, thermal Enhanced language learning tool), was designed specifically for babies within the age range of 6-12 months; Figure 1 [27]. RAVE consists of two agents: a virtual human (provides language and socially contingent interactions) and an embodied robot (provides socially engaging physical cues to babies and directs babies' attention to the virtual human). Of note, the use of a virtual human on a screen provides the benefit of having an expressive agent (both in facial expressions and posture) that can produce a

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 2019, Montreal, Canada

https://doi.org/doi

natural signed language (American Sign Language, ASL) as linguistic input. RAVE brings together science from multiple disciplines to explore the potential for technologies such as functional Near Infrared Spectroscopy (fNIRS) brain imaging that measures the baby's higher cognition, thermal IR imaging that measures the baby's emotional engagement, robotics, and virtual humans in an attempt to positively influence the human learning process (Figure 1). Building on fNIRS studies of infant brains and language processing across infancy, one especially unique feature of RAVE stems from the specific linguistic nature of the avatar's language, which contains the phonetic-syllabic rhythmic temporal patterns that precisely match the human infant brain's peaked sensitivity to these language patterns during this critical period of human language learning [19–22].

While the research science challenges of building a RAVE prototype are complete, the next-step challenge is to evaluate does it work? Can an artificial agent-human language learning system, like RAVE - designed for young infants - be used to facilitate babies' language learning? In this paper, we perform an evaluation of the RAVE system with human babies. Our focus is on the baby's interaction with the virtual human avatar, which is providing multiple kinds of social and linguistic behaviors, and in multiple conversational modes. We asked the following main research questions:

(1) Do babies attend to the avatar and respond to its communicative behaviors?
(2) Can babies with little or no exposure to ASL distinguish among the avatar's different conversational roles (or modes: Linguistic Nursery Rhymes; Social Gestures; Idle/nonlinguistic postures; 3rd person observer)? An important question is whether young babies are able to distinguish among different kinds of avatar behaviors in the first place (particularly as they appear on a flat TV monitor), and if so, how they react to different avatar roles?
(3) Can an avatar stimulate babies' production of socially contingent responses, and crucially, nascent language responses?
(4) What, if anything, is the impact of the presence and/or absence of parents' participation in the conversational interaction? Is intervention from parents in the conversational interaction beneficial for the baby, and how does parental interaction impact the conversational exchange?

Below we report the results from an experimental study using the RAVE system in order to evaluate the system's performance regarding the above questions, with the ultimate goal of evaluating the potential for AI/Avatar systems to facilitate language learning in young babies.

2 BACKGROUND AND MOTIVATION

Language is the principal system of expression and communication for humans and arguably the most prominent cognitive and cultural tool that distinguishes human beings from other species. Acquiring language commences from birth aided by multiple factors, including brain-based sensitivities to aspects of the specific rhythmic patterning of human languages, observation, and engagement in social interactions with the outside world [3]. Language exposure plays an important role in infants' early development of linguistic abilities. Ages 6-12 months is widely recognized as a critical developmental period for language [19, 22]. It is during this period that babies acquire essential phonetic-syllabic segments unique to their native language, which make possible their ability to acquire their native language's vocabulary, discern their language's distributional and syntactic regularities, and crucially, to engage in letter-to-phonetic segment mapping in early successful reading [13, 22]. In early life, select brain sites participate in early human language learning (such as the Planum Temporale in the Superior Temporal Gyrus), which are sensitive to specific rhythmic temporal patterns at the nucleus of phonological structure found in all world languages (spoken and signed) [19, 24]. Exposure to these patterns is crucial for the development of this brain sites and systems to support later healthy language, phonological, reading, and cognitive growth [22]. Children deprived of this early exposure specifically during the ages of 6 to 12 months may face dire consequences such as delays in cognitive, linguistic, reading, and social skills which may last for years [26, 30] with accompanying devastating lifelong impact of reading and academic success [22].

Intriguingly, the same developmental brain sensitivities to the rhythmic temporal patterns of human language phonology also exist in deaf babies learning a natural signed language, and it develops on the identical maturational time table as hearing babies [20, 21, 23]. This universal brain sensitivity enables young sign-exposed babies the early life language input that, in turn, permits them to build a sign phonological system vital to letter-to-sign-phonetic segment mapping in successful reading acquisition [22]. To be sure, all babies who miss exposure to the patterns of their natural language in early life (be it a signed or a spoken language) are rendered at risk for language and reading delays spanning life [19, 22].

Given that 91.7% of young deaf babies are born to non-signing families (hearing) [6], in these families, quickly learning a new signed language can become a challenge for the parents. There are some speech-based interventions such as cochlear implants designed to make available spoken language to the young deaf baby [7, 33]. However, most of these tools cannot be deployed until the ages of 18-24 months. While efforts have begun to implant children at younger ages (from ~8 months), precise adjustments, tuning of the device, as well as intensive speech training, still typically begins after ages 18-24 months and proceeds for months into years thereafter [16]. Thus, this is well past the early critical period for learning phonological units, phonological segmentation, categorization and mapping, and sequencing distributions - all vital to optimal, healthy language learning and reading. As such, there is a pressing opportunity for AI technology that can provide signed language input in the critical period of 6-12 months.

3 THE RAVE SYSTEM

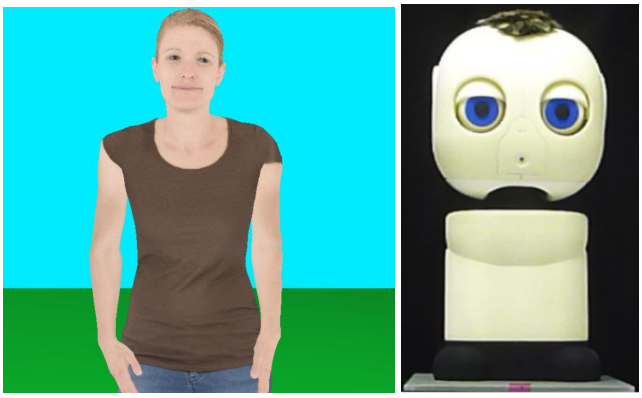
The RAVE system includes two behavioral agents (a physical robot and a virtual human avatar on a screen) that can provide visual behaviors, as well as several sensor devices: an eye-tracker, thermal camera, and an interface for indicating communicative baby behaviors, as seen in Figure 1. Detailed description of the system's constituent components and dialogue algorithms are presented in [15, 27], respectively. Here we summarize and briefly identify the deployment of this AI system to motivate our experimental design.

To that end, we review the components, system's architecture, and briefly describe the behavior selection procedure.

3.1 Agents

The avatar (Figure 2a) provides the linguistic stimuli to the baby. It was built using a real-time character animation system [28], and facial scans from a Light Stage [2]. Avatar behaviors were built by motion capturing a real human deaf native signer of American Sign Language.

The robot (Figure 2b) is based on the open-source Maki platform from Hello Robot [17]. The main purpose of the Robot is to gain the baby's attention and to shift the baby's directional gaze to the avatar. Prior research focusing on the robot component of RAVE demonstrated that the robot functionally achieved this outcome [27]. The robot has an articulated head (pan left/right, tilt up/down), articulated eyes (pan left/right, tilt up/down) and eyelids (open/close). Greater detail about the robot design and impact are presented in [27].



(a) Avatar (b) Robot

Figure 2: Agents of the System

3.2 Perceptual modules

Multiple real-time sensory inputs were used to assess a baby's state of engagement (i.e., attentional, emotional/arousal) to facilitate a socially contingent interaction.

(1) Eye gaze is used as a measure of behavioral response of attention. The baby's eye gaze is categorized as either looking at the Robot, looking at the avatar, looking somewhere in between them, or directed to something else. A Tobii Pro X3-120 [32] was used to recognize the baby's eye gaze at the rate of 120 Hz.
(2) The use of thermal Infrared (IR) imaging, facilitates monitoring the baby's changes in emotional/arousal and attentional engagement as indicated by their Autonomic Nervous System (ANA) responses; i.e., parasympathetic and sympathetic [8, 31] and it is used as a trigger as to when the agents should provide linguistic stimuli to the baby.
(3) A human observer interface was used to capture the communicative and social behaviors of the baby. This feedback from the baby was used as an additional input to our system's dialog manager.

3.3 Agent Behavior Selection

The dialogue management module is constantly updating its internal state based on the sensory input signals as well as the feedback/callback signals from agents. A rule based decision system is used to output signals that are sent to the avatar and the Robot. Detailed explanation about this system is presented in [15].

The design of social contingency for the avatar is built from input using multiple perceptual modules of the system rather than having a fixed protocol. In other words, the avatar would adjust its behavior according to the babies' behavioral responses and attentional/emotional engagement in order to maintain a socially contingent interaction and would provide linguistic input to the baby upon seeing engagement from the baby and proof of its attention (via the triggering from the thermal IR imaging input) [15].

4 AVATAR BEHAVIORS

For the purposes of evaluating the ability of the system to engage in socially contingent interaction with the baby, we focus the analysis on the avatar's different conversational modes, including categories for noncommunicative behavior, social dyadic and triadic (including the robot) behaviors, and those that contain developmentally appropriate linguistic features. The categories used are as follows:

(1) **Idle behaviors ("Idle")** are nonlinguistic/nonsigning, and non socially communicating neutral bodily postures, e.g., arms at side with typical slight body shifting). This behavior typically occurred when the robot has the floor and is engaging with the baby, and avatar is looking at the robot or the baby as a 3rd-party conversationalist.
(2) **Nursery Rhymes ("NR")** are linguistic stimuli such as the "BOAT-ON-WAVE" nursery rhyme in ASL, crucially, with specific rhythmic temporal patterns at the core of all languages' linguistic phonological structure. [1] While the ASL NR is unique to the ASL language and Deaf culture, a rough semantic neighbor in the English language would be "Row-Row-Row-Your Boat" a simple repetitive rhythmic rhyme with approximate versions in many languages around the world.)
(3) **Social Gestures ("Social")** include universal social routines (e.g., BYE-BYE, HI), conversational fillers (e.g., Affirmative Head Nod), and/or short lexical phrases such as YES! or THAT (i.e., English's "right")!
(4) **3-Way behaviors ("3-Way")** are avatar's communicative interactions that were directed to both the baby and the robot, such as "LOOK-AT-ME" (grammatically inflected in the grammar of ASL to include both the baby in second person role and the robot in third person role.

4.1 Design of Nursery Rhyme Behaviors

Linguistic Patterns provide the vital linguistic stimuli for the baby. Nursery Rhymes were constructed with the identical rhythmic

[1] The formal linguistic notation of natural signed languages, such as ASL, uses glosses showing approximate English translations in capital letters.

(a) (b) (c) (d)

Figure 3: Frames of Avatar doing the BOAT Nursery Rhyme.
The four frames were selected from a fluid videoclip of signing as produced by the avatar. Each of the four frames represents a silent sign-phonetic-syllabic contrastive unit as produced with the hands in the American Sign Language visual nursery rhyme, called "BOAT-ON-WAVE" In formal linguistic analyses, these contrastive phonetic-syllabic units are notated as follows (from left to right): 3a /B/+low; 3b /B/+modulation+high; 3c /5/+modulation+high; 3d /5/+modulation+low). The phonetic-syllabic units are shown here in the order of their sequential appearance when produced by the avatar. As in world languages, these phonetic-syllabic linguistic units are not produced in isolation (for example, as a list). Instead, they are bound into fluid movements that form rule-governed, grammatical clausal, phrasal, and syntactic constructions in natural language, here ASL. These phonetic-syllabic contrastive units are produced in ASL with the same maximally-contrasting rhythmic temporal patterning of the phonological level of organization found in languages of the world.

temporal patterning that matched the infant brain's specific neural sensitivity to that rhythmic temporal patterning [20, 22]. All Nursery Rhymes were built with the maximally-contrasting rhythmic temporal patterning in 1.5 Hz alternations [20, 21]. Specific phonetic-syllabic contrasts that human infants first begin to perceive and produce in language development (ages 6-12 months) were used. These include 3 maximally-contrasting phonetic hand primes in ASL: /5/, /B/, /G/ with contrastive transitions /B/⇒/5/, /5/⇒/F/, /G/⇒/F/, plus allophonic variants. Below we provide some examples of the Nursery Rhymes as per formal analyses in the formal discipline of Linguistics analyses for American Sign Language, which had baby-appropriate lexical meaning with their respected action patterned sequences:

- **BOAT[2] (Phonetic-Syllabic units /B/, /5/)**
  (1) BOAT (/B/, double bounce=noun; palms in/+ low center)
  (2) BOAT-on-WATER (/B/+modulation, palms in/+ high center)
  (3) WAVE (ROLLING) (/5/+SAME modulation, palms out/+ high center)
  (4) WAVE (ROLLING) (/5/+SAME modulation, palms down/+ low center)
- **PIG (Phonetic-Syllabic unit: /5/)**
  (1) PIG (/5/, Chin)
  (2) PET (/5/, called "center space" in Linguistic sign notation)
  (3) HAPPY (/5/ + double-handed, Chest)
- **FISH (Phonetic-Syllabic unit: /B/ (allophonic))**
  (1) FISH (/B/, "center space")
  (2) FINS (/B/+double-handed, Head)
  (3) SWIMS (away) (/B/, Cross-Body)
- **CAT (Phonetic-Syllabic units: /5/; /G/allophonic; /BENT5/; and /F/**
  (1) Grandma has red cat [/5/⇒/G/] and [/G/⇒/F/];
  (2) Grandma has white cat [/5/⇒/BENT5/] and [/BENT5/⇒/F/]

5 EXPERIMENT PROTOCOL

In order to address the research questions about the impact of the avatar behaviors on human babies, we designed an experiment whereupon babies interacted with the system in an experimentally controlled setting. While previous investigations were conducted with over 40 babies focusing on this AI/RAVE system's functionality [15, 27], the present study provides a first-time evaluation focusing specifically on the babies: what behaviors did the baby produce in relation to the avatar's behaviors? Is there a principled and predictable relationship? Is there evidence that the babies' behaviors are influenced and/or facilitated by the avatar's behaviors? 4 babies (ages approximately 6-13 months) participated in an intensive case study with a focus on the nature of the babies' behaviors relative to the avatar's behaviors (spanning linguistic and nonlinguistic behaviors while in multiple conversational roles).

As a unique design feature of our avatar's behaviors, we built into its behavioral repertoire the production of patterns that are universal to all infants' brain sensitivity to the linguistic rhythmic temporal patterning underlying phonological organization in language [19]. The strong hypothesis here is that it is the linguistic patterning that is key to the avatar's productions, not the modality. Specifically, we hypothesized that if we had correctly hit on just the right temporal patterning in the avatar's productions, then all babies would be engaged by the avatar's language productions over other social and communicative conversational modes, even if the language patterning was silent and on the hands in ASL - indeed, even in babies who were never exposed to a signed language. Moreover, if correct, we further hypothesized that we would also observe

[2]Sequence of frames of this Nursery Rhyme is depicted in Figure 3.

a capacity in the babies to discriminate among the avatar's multiple categories of behavior (multiple conversational modes), even without understanding the semantic meaning or content of a signed language. To be sure, a finding of this sort would be commensurate with the discipline's widely observed findings that all human babies at approximately 6 months are similarly attracted to the patterns in spoken French, or spoken Spanish, or spoken Swahili, and so on and so forth. This would corroborate the now-classic studies in early infant language processing that demonstrates their ability to discriminate categorically among classes of speech sounds in different languages based on their contrastive patterning (peaked between ages 6-12 months). However, what would be most remarkable here is that we might find this to be true also with a human language in a dramatically different modality, a signed language. Therefore, as the strongest test of this unique language design feature of our avatar, we experimentally tested both non-sign-exposed (n=3) as well as sign-exposed (n=1) babies.

Babies were seated on their parent's lap facing the system (robot and avatar on the screen), on which the avatar would produce its variety of conversational modes. Multiple cameras are used to record the baby (and the parent) from different angles. Each baby's experimental session lasted until the baby became distracted or entered a fussy state in which case we immediately ceased the session. On average, babies stayed engaged with the system for 4 minutes, an intriguing observation in itself for such young babies. The experiment consisted of several steps: upon arrival, the baby and the parent were greeted and introduced to (and greeted by) the robot and then the avatar. This introduction period has been proved to be useful by Meltzoff et al. [14]. Next, a calibration process (a technical requirement of the thermal IR Imaging and Tobii eye tracking systems), followed by the Interaction Session. All experimental sessions were video-recorded for subsequent transcription, coding, and reliability checks by trained experts in child development and linguistics. Figure 4 shows the physical deployment of the experimental setup.

Figure 4: Experimental Setup (Side View)

5.1 Familiarization

To make the baby feel comfortable and involved in this multiparty socially contingent interactions and also to introduce the agents as conversational partners, we begin the experiment with a familiarization episode between the baby, robot and avatar. This process is done with the help of an experimental assistant who interacts with the agents. At the beginning, the assistant talks as well as signing to the robot to wake him up. The robot wakes up, lifts his head, blinks, sees the baby and nods as an acknowledgment of baby's presence. Then it turns toward the avatar. Avatar sees the robot, turns to him, nods, then turns back to baby and waves to the baby. Avatar takes the floor and signs HELLO and GOOD MORNING to the baby to begin the interaction. At this point the assistant signs GOODBYE to the baby and then to the agents, and departs from the experiment room, leaving the baby to interact with the system.

5.2 Interaction Session

The avatar's socially contingent interaction session with the baby began after the assistant left the room (Condition 1). At approximately 2.5 minutes into the experimental session, parents are permitted to interact as per their natural inclination (Condition 2). Throughout the experiment, parents wore sunglasses which was an intentional design feature meant to block the technology from recording eye-tracking artifacts from the parents' eyes. Note that none of the perceptual components were monitoring the parent, so none of the agent's behaviors were contingent directly on the parent [15].

6 RESEARCH QUESTIONS AND EVALUATION METRICS

Scassellati et al. [27] and Nasihati Gilani et al. [15] report observations of several kinds of baby's spontaneous behavioral responses to the Avatar conversational modes. We categorize these as follows:

(1) **Linguistic Responses ("ling")** include manual babbling, the production of manual proto sign-phonetic units, proto-signs, and imitations of signs (i.e., the baby imitates or copies what it sees the Avatar is producing);
(2) **Social/Gestural Responses ("S/G")** include pointing, waving, clapping hands or attempts to copy the agents' behaviors, or social referencing to the parent;
(3) **Sustained Visual Attention ("SVA")** indicates the baby being visually transfixed on the agents for atypically extended periods for infants, defined as greater than one second for this study.

Note that these categories are not mutually exclusive. A baby can exhibit SVA, that is be visually transfixed on the avatar and simultaneously be producing social/gestural responses or linguistic responses. Producing visually transfixed attention, social gestures and especially linguistic behaviors are an indication that system is successful at soliciting babies' interaction. Frequency analyses of the baby's behaviors throughout the experiments provided us with a good insight of the babies' behavioral pattern.

We can now operationalize the main research questions raised in section 1. Regarding the first question (do babies attend to the avatar and respond to its communicative behaviors?); one possibility is that babies do not see the avatar, or the agents collectively, as interesting social interlocutors or respond to them at all. Another possible outcome is that the infants may enter an agitated mode upon confronting an unknown (or "strange") situation such as the

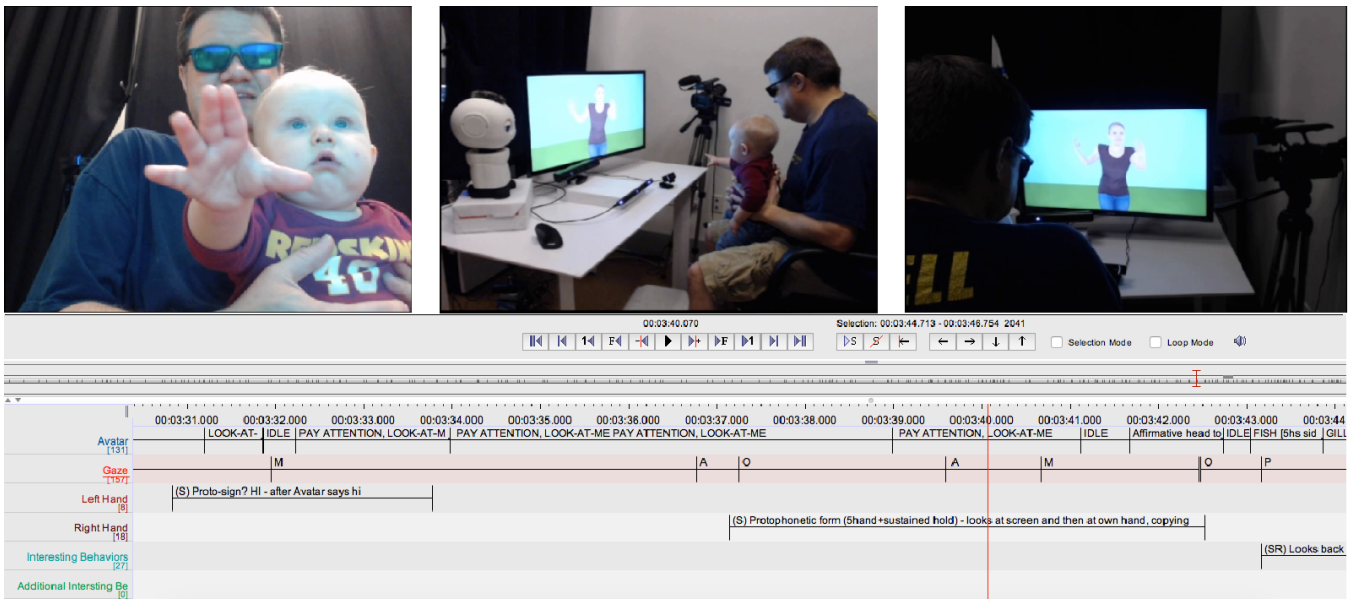


Figure 5: Annotation example using ELAN tool

RAVE system [5]. We use the percentage of baby's responses to the avatar as a metric to evaluate the overall system's impact and performance in terms of engaging the babies.

The second question asked whether babies can differentiate among the different avatar conversational modes. Here, we examine whether there is any difference in baby's response to the avatar's different conversational modes. The intriguing question is this: Are babies able to distinguish among the different avatar's behaviors even though it's unlikely that these young babies understand the semantic content of the ASL language productions (i.e., vocabulary meanings, syntax, etc.)? And, if yes, why? (How might it be possible?)

The third question we asked is of particular scientific interest concerning the mechanisms that drive early language learning: does the avatar's specifically linguistic productions garner the baby's attention, and, in particular, does the avatar's linguistic productions garner linguistic responses from the babies? We hypothesized that if the babies were able to differentiate the avatar's linguistic input from other types of avatar behaviors, then babies would react with more linguistic content when the avatar was in this category, as compared to when the avatar was in its other conversational roles.

Finally, the fourth intriguing question concerns whether having the parent intervene in the conversational interaction is beneficial in terms of facilitating the system's overall language learning goals, or would it have an adverse effect? Perhaps babies would feel more comfortable when they find themselves in a familiar and natural situation in which their parent is part of the interaction and acknowledging their social referencing other than standing still and not reacting to any of their behaviors (which is definitely not a routine for parents). On the other hand, the intervention from the parent might be distracting for the baby and steal the attention from the avatar; as a result, babies may turn to parents for interaction instead of engaging with the system. The first metric to assess this is the overall response rate across conditions. Furthermore, studying the distribution of baby responses across conditions would give us detailed insight on the parent's impact on the social/conversational interaction.

7 EVALUATION AND RESULTS

The video-recorded socially contingent interaction sessions were coded for conversational turns. All coding was done by trained experts in the field of child development and linguistics with scrupulous reliability checks. ELAN tool [1] was used for annotating the baby's behavior as well as marking the times of avatar and robot's behaviors. ELAN is a professional tool to manually and semi-automatically annotate and transcribe audio or video recordings and it has a tier-based data model that supports multi-level, multi-participant annotation of time-based media. A screenshot of the tool along with different tiers is shown in Figure 5.

Here, we present the results of our analyses in two parts. First, we show the interactions between the baby and the avatar, the babies' specific categories of spontaneous behavioral responses, and their relations and dependencies on one another. Second, we show the corresponding analyses regarding the parents, and the impact of parent's intervention on baby's behaviors toward the system.

7.1 Baby and Avatar

In answer to questions 1 & 2 above (do babies attend to the Avatar?; do they differentiate among its conversational modes?) our analyses revealed three discrete categories of behavioral responses that the babies produced to the Avatar (linguistic; sustained visual attention; social/gestural responses), demonstrating the intriguing findings that the babies' indeed attended to the Avatar, and produced differential responses to to it. Next, a frequency analysis of responses to

avatar behaviors was conducted. Figure 6 shows a Venn diagram of the four discrete categories of baby behavioral responses to the avatar as well as the relative frequency of each category of babies' behavioral responses to the Avatar during the Experiment. Note the overlapping portions show cases where the baby responded in more than one way to the same avatar behavior. We can see that the babies' transfixed sustained visual attention (SVA) constitutes the biggest portion of babies' behavioral responses to the Avatar (48% overall).

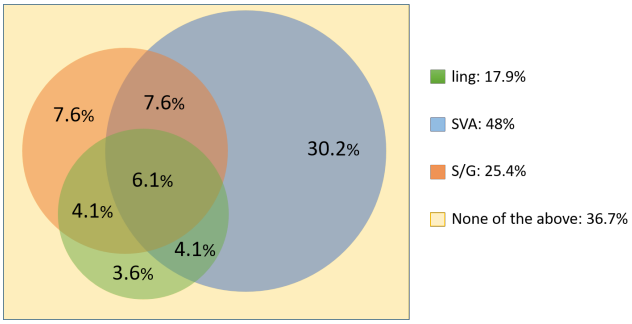


Figure 6: Frequency of Baby's Behaviors

Following from questions 1, 2 and 3 (does the Avatar's linguistic behavior impact linguistic productions in the baby?), we studied the relationship between the Avatar and baby's behaviors. Overall, babies responded to more than 60% of Avatar's behaviors ($M = 61.8, SD = 6.9$). As would be predicted if the babies were attending to the Avatar's different conversational modes, the babies' responses were not equally distributed across different types of Avatar's behaviors. Of crucial scientific significance, babies' response rate appeared to be related to the Avatar's behaviors. Figure 7 shows the baby's response rates to different types of Avatar's behaviors. Regarding the babies' linguistic productions, as a first step in our analyses, we observed that the babies produced their greatest percentage of spontaneous responses to the Avatar when the avatar was producing Linguistic Nursery Rhymes: babies produced spontaneous behavioral responses to 85% of the Avatar's Linguistic Nursery Rhymes, 84% of the 3-Way conversational turns, 75% of the Avatar productions when in Social Gesturing conversational turn, but only 37% of times when the Avatar was idle. The distribution of avatar behaviors was also not uniform: 13% of the avatar's behaviors were NR, 13% 3-way, 36% were social, and the remaining 38% were Idle. The babies' responses to the Avatar's actions (producing Linguistic Nursery Rhymes, social/gestural behaviors or 3-way robot-avatar-baby behaviors) were significantly more compared to when the avatar was in its idle mode ($t = 3.35, p = 0.01$). Thus, here, as above, the babies do appear to attend to and to respond to the Avatar's different conversational modes, with the babies' greatest percentage of responses being when the Avatar was producing Linguistic Nursery Rhymes.

Further to question 3 above, we investigate the relationship between baby and avatar behaviors, whereby a frequency analysis was conducted of the different baby behaviors in response to the avatar's behavior. Figure 8 shows the rate of each baby behavior in response to each category of avatar's production. Note, here we

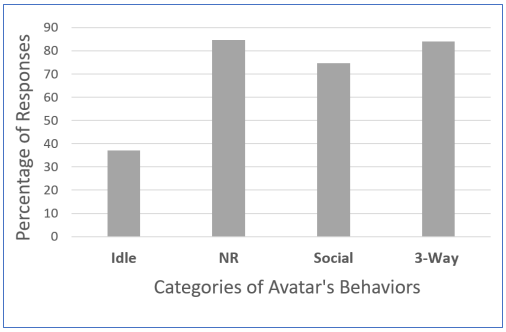


Figure 7: Percentage of Babies' Responses to Different Avatar Behaviors

analyzed the immediate responses from the babies that occurred just after or at the same time as the avatar behaviors. However, as would be conversationally appropriate, babies did produce perseverative responses toward the avatar, which were largely linguistic in nature; for example, babies would begin to copy the avatar's sign after a few seconds delay, even though the avatar might have segued to its next conversational role. Note that the bars in each category in this Figure do not necessarily need to add up to 1, because sometimes the baby responds with multiple response types, as shown in Figure 6.

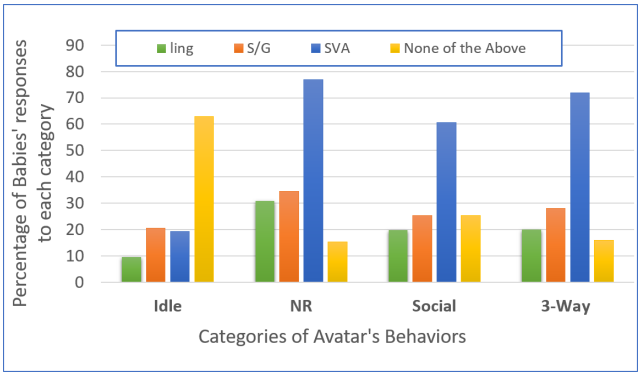


Figure 8: Babies' Categorical Responses to Different Avatar behaviors

Of particular theoretical importance regarding the present scientific questions, the babies responded differently when the avatar was in the Linguistic Nursery Rhyme conversational mode versus when the avatar was in other conversational modes (Social Gestures, Idle, 3-Way). The babies produced the largest percentage of linguistic responses to the avatar's Linguistic Nursery Rhymes (31% to Nursery Rhymes vs 10% to Idle, 19% to Social Gestures, and 20% to 3-Way). Further, the babies' responses to the avatar's Linguistic Nursery Rhymes (over the avatar's other conversational turn types) involved them to be largely riveted into a state of fixed and Sustained Visual Attention (77%). Of theoretical significance, there appears to have been a principled relationship between the avatar's socially contingent communicative turn types and the babies' specific responses. This relationship implies that the avatar was indeed having a linguistic impact on the baby.

7.2 Parent's Intervention

To address question 4 (impact of +/- parental intervention), we first focused on the role of the parent in the conversational interaction. We analyzed the different baby behaviors across the two conditions. Interestingly, babies responded to 80% of avatar's behaviors in Condition 1 versus 60% in Condition 2 ($t = 2.22, p < 0.05$). This decrease is mainly due to a significant decrease in the babies sustained visual attention, SVA ($t = 4.3, p < 0.005$). This finding makes sense since in condition 2, parents were acknowledging and interacting with the baby, whereupon babies would naturally look more at the parent thereby exhibiting fewer instances of sustained attention toward the avatar.

We also analyzed the distribution of baby responses across the two conditions (Figure 9). Here we see a significant increase in the percentage of babies' linguistic behaviors across Condition 1 vs Condition 2 ($t = 2.4, p < 0.05$). This is a very interesting finding, as it indicates that parents' interactions when involving both their baby and the RAVE system (combined) may have the potential to augment the language learning impact of RAVE. Apart from parental impact, the present pattern of change from Condition 1 to Condition 2 may imply that the infant is evidencing aspects of learning (to be further explored).

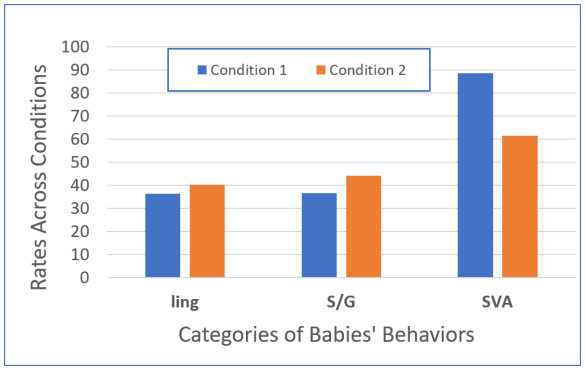


Figure 9: Relative Frequency of Baby Response Types given Absence (Condition 1) or Presence (Condition 2) of Parental Involvement

8 CONCLUSION AND FUTURE WORK

The driving theoretical question, and novelty, of the present paper was to understand whether an artificial agent (the ASL signing Avatar) had the potential to facilitate language learning in young babies.

To address this, we studied the impact that a signing avatar agent (with its richly varied communicative and socially contingent conversational modes) had on young babies' spontaneous behavioral responses, in particular, we asked whether the avatar's linguistic productions in signed language would spontaneously trigger linguistic responses from the babies (be they babies exposed to a signed language or babies exposed to a spoken language). We were especially interested if a very young baby would even detect the avatar's different conversational modes, as the avatar was projected onto a flat television screen.

We indeed found that babies spontaneously distinguished among avatar conversational modes. Babies produced different categories of behavioral responses to the avatar, and, further, their different behavioral responses were socially contingent (related to) the avatar's different conversational modes. To be sure, the results indicate that the babies were indeed able to detect the avatar's different conversational modes even though all appeared on a flat screen.

One surprising and unexpected finding was that babies produced the greatest percentage of linguistic responses to the Avatar's linguistic Nursery Rhymes versus other avatar conversational modes. The babies' linguistic responses included linguistic ASL-related productions (spanning manual babbling, production of proto-sign phonetic units and protosigns, linguistic sign-phonetic and sign imitations). Here, the most intriguing question is why? How was it possible that the young baby was able to detect the avatar's different conversational modes, producing the greatest linguistic responses to the avatar's linguistic productions, especially when most of the babies did not understand American Sign Language and thus could not possibly have been understanding the meanings of the language that they were observing.

Herein lies one of the most powerful and scientifically revealing findings of the present study concerning the nature of the brain-based mechanisms that govern human language acquisition. Rather than being attracted to the meanings of the language before them (which, again, most did not understand), we hypothesized that all babies (deaf and hearing) were differentiating among the avatar's conversational modes based on differences in their +/- relation to the rhythmicity of human language phonetic-syllabic (phonological) structure. Because ASL is a real language and possesses the phonological structure universal to all world languages, and because the avatar was producing the rhythmic temporal patterns that underlie phonological structure in ASL, babies demonstrated riveted attention to this category of avatar productions. This finding is much like those showing that babies demonstrate riveted attention to the phonological patterns in their native language as well as in a foreign (non-native) language over other patterns of acoustic stimuli [10, 13, 22]. These findings provide support for the hypothesis that infants are born with a sensitivity to specific rhythmic temporal patterns in language and that the avatar had hit squarely on those patterns, a finding to which we will devote future experimental work.

The present findings also suggest that the AI-Avatar dialogue system had achieved a level of verisimilitude to social contingency found in natural parent-baby discourse. Indeed, the babies' interactive engagement with this artificial avatar agent provides remarkable evidence that social contingency is a vital component of healthy language learning. Beyond the importance of social interactions, the role of social contingency in early human language acquisition will also be pursued in our future work along with work analyzing the robot's role in the system.

Nonetheless, all of the babies' appeared to be captivated by the avatar, and exhibited spontaneous engagement with the avatar, which was powerfully observed and which occurred even though the avatar is an inanimate artificial agent on a flat TV monitor. In conclusion, the present work provides a novel demonstration of the potential for avatars to facilitate language learning in young babies.

REFERENCES

[1] Hennie Brugman, Albert Russel, and Xd Nijmegen. 2004. Annotating Multi-media/Multi-modal Resources with ELAN.. In *LREC*.

[2] Paul Debevec. 2012. The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia* 2, 4 (2012).

[3] Amy Sue Finn. 2010. *The sensitive period for language acquisition: The role of age related differences in cognitive and neural function.* University of California, Berkeley.

[4] Ewa M Golonka, Anita R Bowles, Victor M Frank, Dorna L Richardson, and Suzanne Freynik. 2014. Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer assisted language learning* 27, 1 (2014), 70–105.

[5] David J Greenberg, Donald Hillman, and Dean Grice. 1973. Infant and stranger variables related to stranger anxiety in the first year of life. *Developmental Psychology* 9, 2 (1973), 207.

[6] P Higgins. 1980. *Outsiders in a hearing world.* SAGE Publishing.

[7] William F House. 1976. Cochlear implants. *Annals of Otology, Rhinology & Laryngology* 85, 3 (1976), 3–3.

[8] Stephanos Ioannou, Vittorio Gallese, and Arcangelo Merla. 2014. Thermal infrared imaging in psychophysiology: potentialities and limits. *Psychophysiology* 51, 10 (2014), 951–963.

[9] Jiyou Jia. 2009. An AI framework to teach English as a foreign language: CSIEC. *AI Magazine* 30, 2 (2009), 59.

[10] Peter W Jusczyk, Derek M Houston, and Mary Newsome. 1999. The beginnings of word segmentation in English-learning infants. *Cognitive psychology* 39, 3-4 (1999), 159–207.

[11] Marina Krcmar. 2011. Word learning in very young children from infant-directed DVDs. *Journal of Communication* 61, 4 (2011), 780–794.

[12] Marina Krcmar, Bernard Grela, and Kirsten Lin. 2007. Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology* 10, 1 (2007), 41–63.

[13] Patricia K Kuhl. 2004. Early language acquisition: cracking the speech code. *Nature reviews neuroscience* 5, 11 (2004), 831.

[14] Andrew N Meltzoff, Rechele Brooks, Aaron P Shon, and Rajesh PN Rao. 2010. "Social" robots are psychological agents for infants: A test of gaze following. *Neural networks* 23, 8-9 (2010), 966–972.

[15] Setareh Nasihati Gilani, David Traum, Arcangelo Merla, Eugenia Hee, Zoey Walker, Barbara Manini, Grady Gallagher, and Laura-Ann Petitto. 2018. Multimodal Dialogue Management for Multiparty Interaction with Infants. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 5–13.

[16] Johanna Grant Nicholas and Ann E Geers. 2007. Will they catch up? The role of age at cochlear implantation in the spoken language development of children with severe to profound hearing loss. *Journal of Speech, Language, and Hearing Research* 50, 4 (2007), 1048–1062.

[17] Tim Payne. 2018. MAKI - A 3D Printable Humanoid Robot. https://www.kickstarter.com/projects/391398742/maki-a-3d-printable-humanoid-robot. (2018).

[18] Laura-Ann Petitto. in press. The Impact of Minimal Language Experience on Children During Sensitive Periods of Brain and Early Language Development: Myths Debunked and New Policy Implications. retrieved from http://petitto.net/wp-content/uploads/2014/04/Petitto_Minimal-Language-Experience_Final_Oct-6-2017.pdf.

[19] Laura-Ann Petitto, Melody S Berens, Ioulia Kovelman, Matt H Dubins, K Jasinska, and M Shalinsky. 2012. The "Perceptual Wedge Hypothesis" as the basis for bilingual babies' phonetic processing advantage: New insights from fNIRS brain imaging. *Brain and language* 121, 2 (2012), 130–143.

[20] Laura Ann Petitto, Siobhan Holowka, Lauren E Sergio, Bronna Levy, and David J Ostry. 2004. Baby hands that move to the rhythm of language: hearing babies acquiring sign languages babble silently on the hands. *Cognition* 93, 1 (2004), 43–73.

[21] Laura Ann Petitto, Siobhan Holowka, Lauren E Sergio, and David Ostry. 2001. Language rhythms in baby hand movements. *Nature* 413, 6851 (2001), 35.

[22] Laura-Ann Petitto, Clifton Langdon, Adam Stone, Diana Andriola, Geo Kartheiser, and Casey Cochran. 2016. Visual sign phonology: Insights into human reading and language from a natural soundless phonology. *Wiley Interdisciplinary Reviews: Cognitive Science* 7, 6 (2016), 366–381.

[23] Laura Ann Petitto and Paula F Marentette. 1991. Babbling in the manual mode: Evidence for the ontogeny of language. *Science* 251, 5000 (1991), 1493–1496.

[24] Laura Ann Petitto, Robert J Zatorre, Kristine Gauna, Erwin James Nikelski, Deanna Dostie, and Alan C Evans. 2000. Speech-like cerebral activity in profoundly deaf people processing signed languages: implications for the neural basis of human language. *Proceedings of the National Academy of Sciences* 97, 25 (2000), 13961–13966.

[25] Rebekah A Richert, Michael B Robb, and Erin I Smith. 2011. Media as social partners: The social nature of young children's learning from screen media. *Child Development* 82, 1 (2011), 82–95.

[26] Jenny R Saffran, Ann Senghas, and John C Trueswell. 2001. The acquisition of language by children. *Proceedings of the National Academy of Sciences* 98, 23 (2001), 12874–12875.

[27] Brian Scassellati, Jake Brawer, Katherine Tsui, Setareh Nasihati Gilani, Melissa Malzkuhn, Barbara Manini, Adam Stone, Geo Kartheiser, Arcangelo Merla, Ari Shapiro, David Traum, and Laura-Ann Petitto. 2018. Teaching Language to Deaf Infants with a Robot and a Virtual Human. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 553.

[28] Ari Shapiro. 2011. Building a character animation system. In *INTERNATIONAL Conference on Motion in Games*. Springer, 98–109.

[29] Glenn Stockwell. 2007. A review of technology choice for teaching language skills and areas in the CALL literature. *ReCALL* 19, 2 (2007), 105–120.

[30] Adam Stone, Laura-Ann Petitto, and Rain Bosworth. 2018. Visual sonority modulates infants' attraction to sign language. *Language Learning and Development* 14, 2 (2018), 130–148.

[31] M Teena and A Manickavasagan. 2014. Thermal infrared imaging. In *Imaging with Electromagnetic Spectrum*. Springer, 147–173.

[32] Tobii Eyetracker 2018. Tobii Pro X3-120. https://www.tobiipro.com/product-listing/tobii-pro-x3-120/. (2018).

[33] Blake S Wilson, Charles C Finley, Dewey T Lawson, Robert D Wolford, Donald K Eddington, and William M Rabinowitz. 1991. Better speech recognition with cochlear implants. *Nature* 352, 6332 (1991), 236–238.